Task Parallelness: Investigating the Difficulty of Two Spoken Narrative Tasks

Chihiro Inoue

CRELLA Research Seminar

July 7th, 2011

Presentation Outline

1. Rationale
2. Summary of Literature
3. Tasks in Question
4. Research Questions & Methodology
5. Results & Discussion
6. Conclusions & Implications
7. Contributions
8. Limitations

1. Rationale (1)

Spoken narrative tasks =

- tasks with a sequence of pictures based on which candidates are asked to orally narrate a story
- Used in some English speaking tests (e.g. TSE, ELSA, Eiken, SST (Japan)); suitable for candidates with relatively lower-proficiency (Fulcher, 2003)

➢ In Japan, the Ministry of Education has launched a large-scale action plan in 2003 for English education reform towards stronger productive skills, especially speaking

1. Rationale (2)

- Equivalent tasks are vital in speaking tests, but evidence of equivalence is not often provided (Weir & Wu, 2006)
- In SLA (task-based research), evidence of equivalence is seldom found before manipulating the design of tasks or conditions of administration (Weir & Wu, 2006)

➔ **How can 'equivalence' or 'parallelness' be established?**

➢ 'Parallel' is narrower than 'equivalent': refers to being designed to be as similar as possible = same instructions, response type, and content. Should yield the same M and SD of the scores (Alderson, Clapham, & Wall, 1995: 96)

2. Summary of Literature (1)

How has 'equivalence' been established in language testing?

(e.g. Shohamy, 1994; O'Loughlin, 2001; Weir & Wu, 2006)

➔**MFRM analysis of scores on performance**

➔ Construct validity (Messick, 1996)

- **Generalizability** (elicited linguistic performance)
- **Substantive aspect** (candidate perceptions)

➔ *A priori analysis* of **task characteristics** & *a posteriori* evidence (i.e. *expected* and *actual* performance elicited)

2. Summary of Literature (2)

How is the evidence operationalized in LT and SLA (TBLT)?

- **MFRM analysis:** use of FACETS
- **Linguistic performance**: syntactic & lexical complexity, accuracy, fluency, idea units
- **Candidate perceptions**: questionnaires, interviews and observations
- **Task characteristics:** factors of "task complexity" (Robinson, 2001; Skehan, 1998): expected syntactic & lexical complexity, topic familiarity, the number of elements, demand for reasoning, prior (background) knowledge

  ➔ Assumption: If these are different, one task is more cognitively difficult than the other

2. Summary of Literature (3)

In sum, it is necessary to collect…

- Ratings (by multiple raters with rating scales)
- Quantified data of syntactic & lexical complexity, accuracy, fluency, idea units ← Validity evidence of such variables (= correlation with ratings)
- Baseline data from NS of English
- Candidate perceptions of task difficulty
- Expert Judgement on task complexity (expected syntactic & lexical complexity, topic familiarity, the number of elements, demand for reasoning, prior (background) knowledge)

3. Tasks in Question (1)

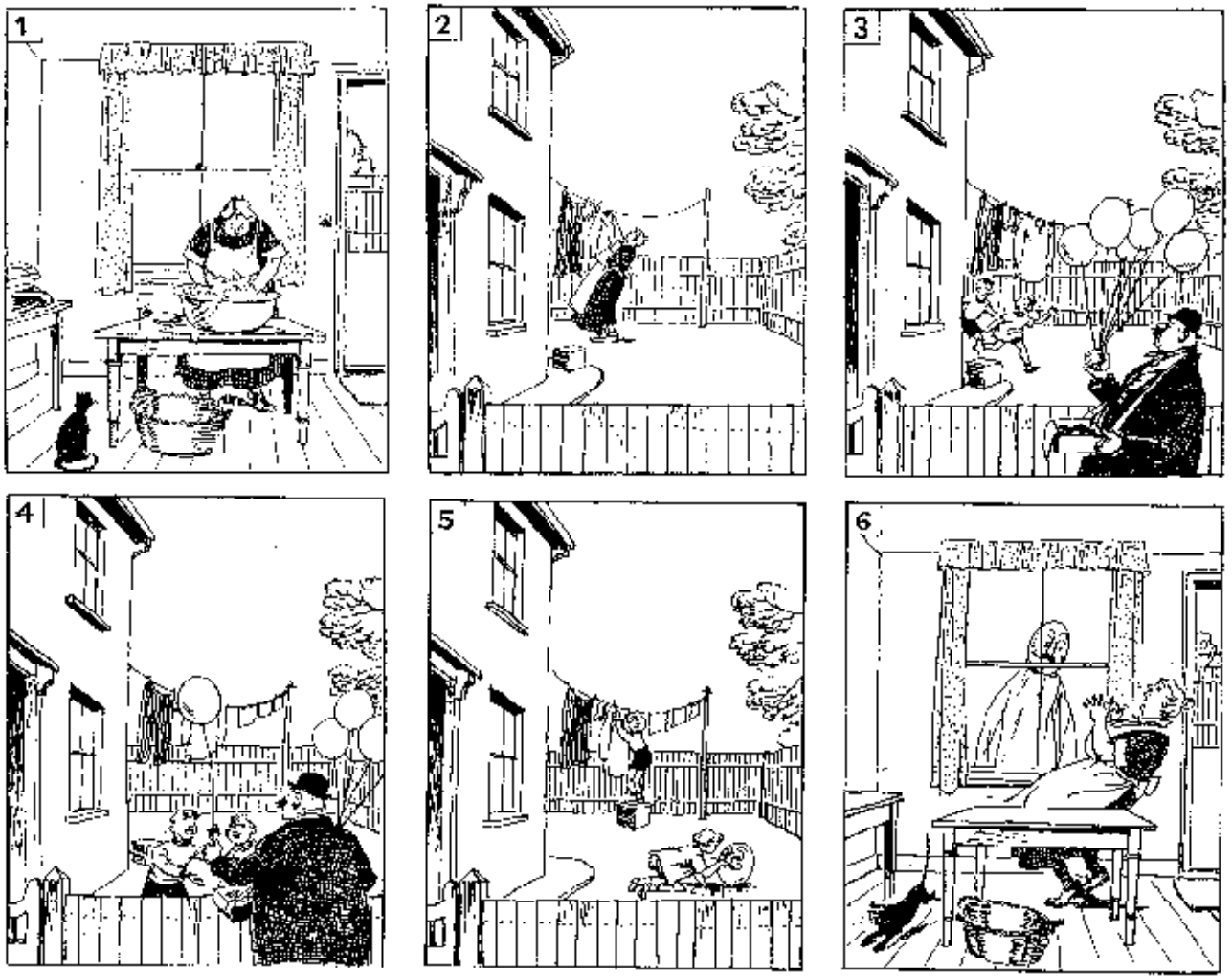
- A pilot study using two supposedly ‘parallel’ tasks from the Standard Speaking Test in Japan
- Tasks with “a conflict in a public place”

→ **NOT** actually parallel in terms of expert judgement and linguistic performance of JS and NS, because of the differences of relationships among characters, prominence of characters & resulting damages.

→ Need for ‘more similar’ tasks

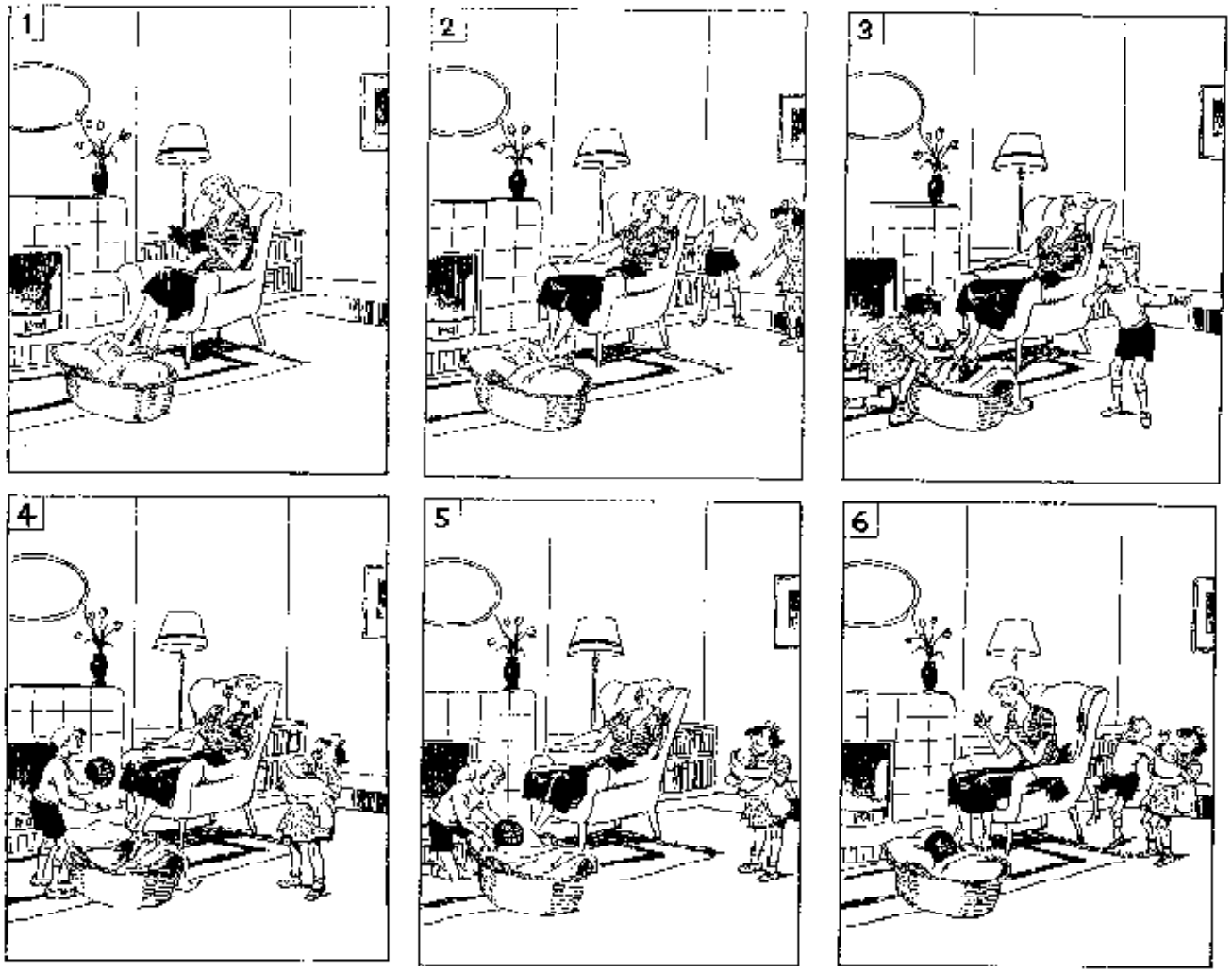
3. Tasks in Question (2) by Hill (1960)

A



3. Tasks in Question (3) by Hill (1960)

B



4. Research Questions & Methodology (1)

RQ1. Is the difficulty of the two tasks the same according to MFRM analysis?

Data:

- 65 Japanese candidates (modern language majors at university)
- 7 raters
- Ratings from Below A1 to C1 based on CEFR Oral Assessment Grid (Range, Fluency, Accuracy, Coherence, Sustained Monologue, and Considered Judgement) on both tasks
- Task difficulty calculated based on Considered Judgement & other 5 rating categories

4. Research Questions & Methodology (2)

RQ2-1. Are the candidates' perceptions of the two tasks the same?

RQ2-2. What about at different levels of proficiency?

→ Related sample *t*-tests on the responses on a 9-point scale questionnaire by Robinson (2001) on perceptions of

task difficulty, nervousness, self-rating of performance, enjoyment, and interest

→ CEFR levels were assigned to each candidate by rounding up his/her fair average calculated by FACETS (to assign CEFR levels from A1 (=1) to C1 (=10) as in Eckes (2009))

4. Research Questions & Methodology (3)

RQ2-3. Do Japanese teachers judge the two tasks to be parallel for the candidates in terms of the relevant task complexity factors?

→ Responses by 2 Japanese teachers on Checklist for Difficulty (Weir & Wu, 2006) (e.g. “The lexical items required are equally familiar to the candidates.”) and in short follow-up interviews

RQ2-4. Do English native speakers perceive the two tasks equally difficult?

→ Responses by 11 NS to the question, “Did you think the two tasks were equally difficult? (If no, why?)” in a short afterwards interview

4. Research Questions & Methodology (4)

RQ3-1. Are the linguistic performances of the two spoken narrative tasks the same in terms of the linguistic performance variables?

RQ3-2. What about at different levels of proficiency (incl. NS)?

→ Related sample *t*-tests (or Wilcoxon signed rank tests) were run:

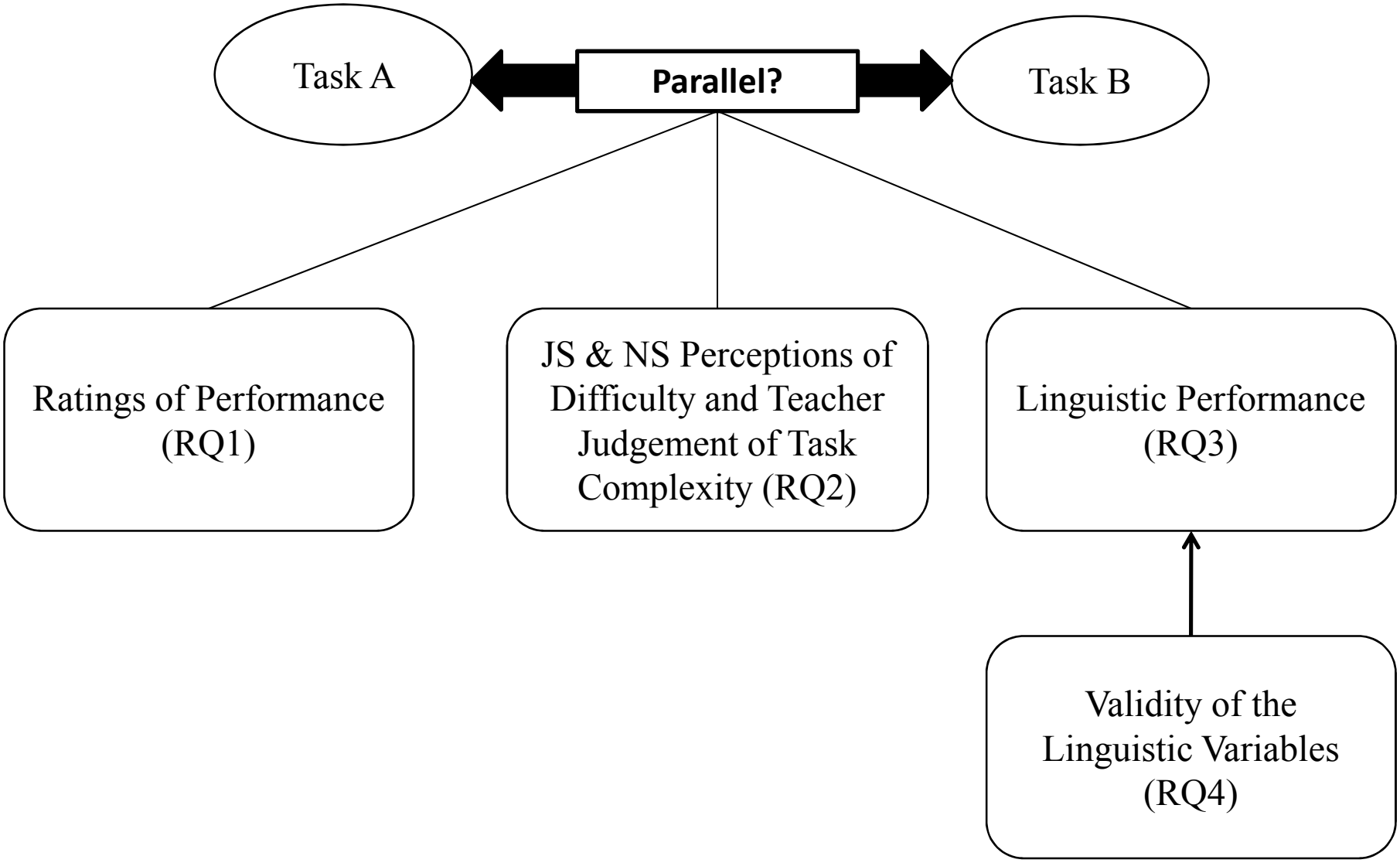
Aspects	**Variables**
Fluency	Speech rate
Accuracy	% of error-free clauses Errors per AS-unit Errors per 100 words
Lexical complexity	D value
Syntactic complexity	AS-unit length Subordinate clauses per AS-unit
Idea units	No. of main idea units No. of minor idea units

4. Research Questions & Methodology (5)

RQ4. How do the linguistic variables correlate with the ratings of spoken narrative performance in the corresponding rating categories?

➔ Pearson's correlation between:

Ratings in	**Variables**
Range	D value (lex. complexity) AS-unit length (synt. complexity) Sub. clauses per AS-unit
Fluency	Speech rate
Accuracy	% of error-free clauses Errors per AS-unit Errors per 100 words
Coherence	Incidence of coordination (Coh-metrix)
Sustained Monologue	No. of main idea units No. of minor idea units


Task A
Parallel?
Task B
Ratings of Performance (RQ1)
JS & NS Perceptions of Difficulty and Teacher Judgement of Task Complexity (RQ2)
Linguistic Performance (RQ3)
Validity of the Linguistic Variables (RQ4)

5. Results & Discussion: RQ1 (1)

Task difficulty calculated by MFRM analysis were:

[Considered Judgement ratings] – with Below A1 to C1

Task A: -0.14 logits

Task B: -0.54 logits = 0.24 points of difference in fair average ratings

[Ratings in 5 ratings categories] – with collapsed levels of Below A2; A2/A2+; B1/B1+; B2/B2+; C1

Task A: 1.66 logits

Task B: 1.14 logits

➔ **Task A was significantly more difficult**

5. Results & Discussion: RQ1 (2)

How 'big' or 'small' is the *small but significant* difference?

> On a golf-course, a hole with 0.24 average strokes more than another hole is considered noticeably more difficult. In your situation, I don't know. 0.1 score-points difference would definitely be "the same". 0.5 score-points difference would definitely be "different". 0.24 is in the gray-area where detailed knowledge of the situation is needed (Linacre, 2011, personal communication).

I would argue that this difference is rather big, considering all the effort to select and make the tasks as parallel as possible. 24 out of 65 candidates would be assigned different (neighboring) levels on Tasks A and B.

5. Results & Discussion: RQs2-1 & 2-2

*Candidate perceptions of their nervousness and self-ratings of performance showed significant order effect.

[RQ2-1] No significant difference between Task A and B on perceived difficulty, enjoyment and interest.

[RQ2-2] No significant difference was found either, however, at B2/B2+ level (*n* = 8), the perceived difficulty was approaching significance (with Task A perceived as more difficult).

5. Results & Discussion: RQs2-3 & 2-4

Expert judgement by 2 Japanese teachers and perceptions by the NS indicated that the two tasks were not parallel.

Task A was more difficult because:

- Time gap between Pictures 5 and 6 which led to insufficiency of details as to how the ghost-like figure was made
- Locations of the room, window, and garden might be difficult to grasp at once
- Lack of washing-related vocabulary of the Japanese candidates (i.e. *washing-line, basin*)

5. Results & Discussion: RQ3

- Less complex (i.e. subordination) and less accurate performance with more main idea units on Task A (RQ3-1)
- Less complex (i.e. subordination), less accurate(?), less fluent (at B1/B1+ level) performance with more main idea units on Task A.
- However, even at A2/A2+ level, significantly less complex performance was elicited (= more subordination was produced on Task B <u>across all levels</u>)

  ➔ questions arose with variables of accuracy and syntactic complexity (subordination)

5. Results & Discussion: RQ4 (1)

Pearson's correlation coefficients for both tasks are shown as (.xxx, .yyy) below. **Significant at .01 level; *Significant at .05 level

Rating Category	Variables	Pearson's Coefficients (A, B)
Range	D value AS-unit length Sub. clauses per AS-unit	.470**, .469** .509**, .282* .265*, .120
Fluency	Speech rate	.806**, .795**
Accuracy	% of error-free clauses Errors per AS-unit Errors per 100 words	.644**, .683** -.652**, -687** -.723**, -.731**
Coherence	Incidence of coordination	-.193, -.122
Sustained Monologue	No. of main idea units No. of minor idea units	.501**, .172 .345**, .375**

5. Results & Discussion: RQ4 (2)

- Highly rated candidates in Range did not necessarily produce longer AS-unit or more subordinate clauses on Task B.

  ➔Examining the transcripts revealed that more subordination was produced at all levels because of the constant presence of the mother and the plot of exchanging the baby with a ball on Task B

- The discrepancy among the results by accuracy variables were due to the spread of errors and difference of denominators
- Two of the main idea units on Task B (out of 9) conveyed redundant content as the other main units, which may have led to candidates not mentioning them.

6. Conclusions & Implications (1)

In summary, Task A and B were **NOT** parallel in terms of:

- Task difficulty by MFRM analysis based on the ratings
- Reports by Japanese teachers and NS
- Syntactic complexity, accuracy and main idea units of the linguistic performance (N = 65)
- Fluency, accuracy, syntactic complexity (subordination) of linguistic performance at B1/B1+ level
- Speech rate, variables of accuracy, D value were valid (i.e. in accordance with ratings)

➔ There is more to ensuring task parallelness than the task complexity factors specify: time gaps between the pictures, sufficiency of details, lexical knowledge and background knowledge (cf. candidate factors)

6. Conclusions & Implications (2)

Questions arise about:

(1) If the assumption that more cognitively difficult tasks elicit more complex language is so generalizable; subordination can be elicited regardless of the complexity of tasks

(2) If measuring the complexity of language by the amount of subordination is appropriate

(3) How to make sure that a task is 'more cognitively difficult' than another, and to confidently conclude that the changes in linguistic performance are attributed to the differences in cognitive difficulty (i.e. task complexity).

7. Contributions & Limitations

(1) Multi-method analysis towards parallelness at task level

(2) Empirical assignment of CEFR levels

(3) Validity study of the linguistic variables

(4) Collecting evidence of task complexity (i.e. cognitive difficulty of tasks)

(5) Collecting NS performance data

(6) Challenging the theories of task complexity

++

(1) A larger sample was desirable (generalizability of findings)

(2) Use of CEFR Assessment Grid might not have been the best choice (lack of correspondence between the descriptors and the variables)

(3) No interview data from the Japanese candidates